2015

Knowledge Enabled Location Prediction of Twitter Users

Revathy Krishnamurthy
Wright State University

Knowledge Enabled Location Prediction of Twitter Users

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

By

REVATHY KRISHNAMURTHY
B.E., University of Pune, 2007

2015
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

February 23, 2015

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY REVATHY KRISHNAMURTHY ENTITLED Knowledge Enabled Location Prediction of Twitter Users BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Amit P. Sheth
Thesis Director

Mateen Rizki
Chair, Department of Computer Science and Engineering

Robert E. Fyffe
Vice President for Research and
Dean of the Graduate School

Committee on
Final Examination

Amit P. Sheth, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Derek Doran, Ph.D.

ABSTRACT

KRISHNAMURTHY, REVATHY. M.S., Department of Computer Science and Engineering, Wright State University, 2015. *Knowledge Enabled Location Prediction of Twitter Users.*

As the popularity of online social networking sites such as Twitter and Facebook continues to rise, the volume of textual content generated on the web is increasing rapidly. The mining of user generated content in social media has proven effective in domains ranging from personalization and recommendation systems to crisis management. These applications stand to be further enhanced by incorporating information about the geo-position of social media users in their analysis.

Due to privacy concerns, users are largely reluctant to share their location information. As a consequence of this, researchers have focussed on automatic inferencing of location information from the contents of a user's tweets. Existing approaches are purely data-driven and require large training data sets of geotagged tweets. Furthermore, these approaches rely solely on social media features or probabilistic language models and fail to capture the underlying semantics of the tweets.

In this thesis, we propose a novel knowledge based approach that does not require any training data. Our approach uses Wikipedia, a crowd sourced knowledge base, to extract entities that are relevant to a location. We refer to these entities as *local entities*. Additionally, we score the relevance of each local entity with respect to the city, using the Wikipedia Hyperlink Graph. We predict the most likely location of the user by matching the scored entities of a city and the entities mentioned by users in their tweets. We evaluate our approach on a publicly available dataset consisting of 5119 Twitter users across continental United States and show comparable accuracy to the state-of-the-art approaches. Our results

demonstrate the ability to pinpoint the location of a Twitter user to a state and a city using Wikipedia, without needing to train a probabilistic model.

Contents

List of Figures

List of Tables

Acknowledgment

My journey through graduate school has been an extraordinary and a fulfilling experience. I would like to take this opportunity to thank everyone who has helped me along the way.

First and foremost, I want to express my sincere gratitude towards my advisor Dr. Amit P. Sheth for his continuous guidance. I am thankful to him for providing me the opportunity to pursue my research interests. His dedication and enthusiasm continues to inspire me every day. I would like to thank Dr Prasad for all the in-depth discussions and his insightful comments. I would also like to thank Dr. Doran for his patience and feedback on this work. I truly appreciate his time and support on giving this thesis its current shape.

I would like to thank Pavan Kapanipathi for guiding me through every stage of my research. I am grateful to him for his kindness and his patience. Thanks also to the entire team of Kno.e.sis. Special thanks to Tonya Davis for always helping me out with the administrative tasks.

This acknowledgement would be incomplete without thanking my family. To my sister Uma, for her endless encouragement and never ending faith in me. I would not have reached here without her cheering me on every step of the way. Last but not the least, I am thankful to my parents Bagyalakshmi and Krishnamurthy, for all their sacrifices and their continued faith in me. This thesis is dedicated to you.

Dedicated to my parents Bagyalakshmi and Krishnamurthy

Introduction

The creation of World Wide Web in 1990 changed the landscape of modern life. The web has grown exponentially since that time. In 1995, less than 1% of the world population had an internet connection. In contrast, approximately 40% of the world population is able to access the internet today. In the initial days, the web was mainly used for displaying information. In comparison, Web 2.0 used technologies that went beyond static pages. The rise of Web 2.0 is characterized by social networking sites, blogs and mashups with focus on participation instead of merely publishing [1].

The rapid growth of social media has changed the way we communicate and express ourselves. Social media's impact across countless facets of society is constantly surprising; for example, Beyonce utilized social media to release her 2013 album in a radical and effective way [2]. The news generated 1.2 million tweets and her album sold 365,000 copies on its first day. Twitter, a micro-blogging website, is widely regarded to be the de facto social media platform [5, 51] with over 255 million users. As of 2014, Twitter users post 277,000 tweets per minute and almost 500 million tweets per day. As shown in Figure 1.1, the average number of tweets posted per day has been increasing at an exponential rate

[1] http://oreilly.com/web2/archive/what-is-web-20.html

[2] http://www.forbes.com/sites/jackiehuba/2013/12/17/beyonce-uses-only-word-of-mouth-to-market-surprise-new-album/

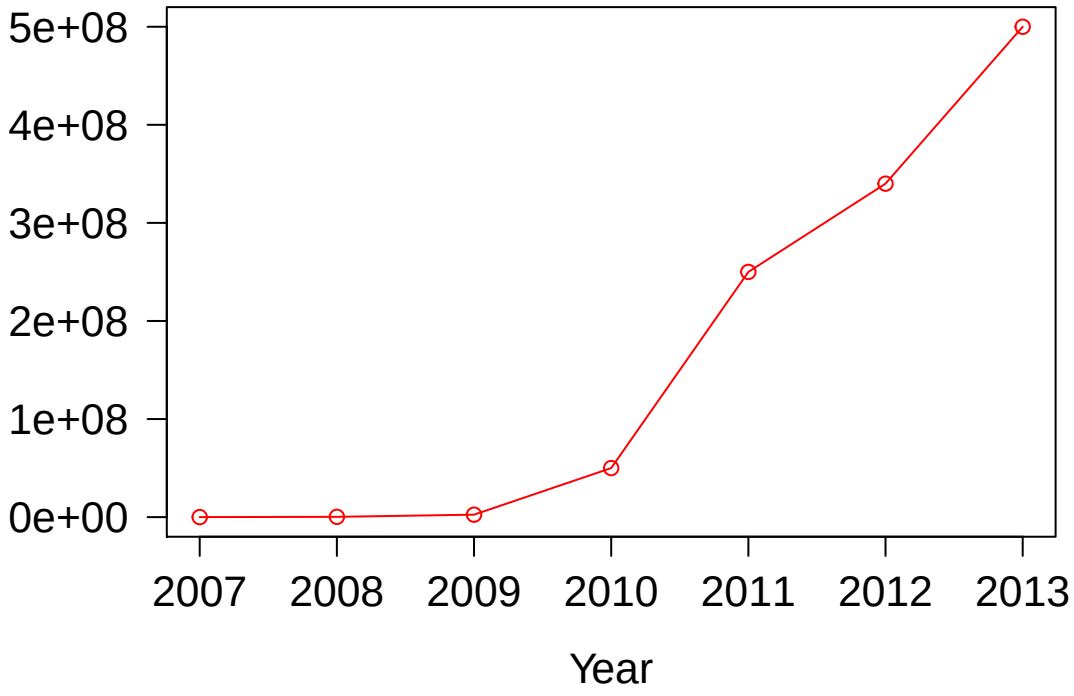

Figure 1.1: Average Tweets per Day

every year. The topic of these tweets may range from what they had for lunch to the economic policies of the United States. The wide range of topics discussed on Twitter have led to the development of applications such as flu outbreak prediction [50], opinion analysis [43], earthquake prediction [49] and smart city projects [3].

An emerging trend has been the geographic footprint of online users [3]. Associating geographic information with social media users can provide value addition to many applications. For example, in the analysis of public opinion on key social issues, location information of users may be used for localizing opinions by geographic regions. Similarly, personalization and recommendation systems may use location information to provide ad-

[3] http://blogs.wsj.com/digits/2014/04/08/the-race-to-locate-twitter-users/

ditional context about a user's immediate surroundings and environments to determine user preferences based on his/her geographic position. Furthermore, location data is central to many applications such as event detection [49] and emergency response systems [34]. Such use cases may specifically include:

1. **Opinion Analysis on Political and Social Issues:** The popularity and wide spread user base of Twitter make it an important platform for public opinion research. American Association for Public Opinion Research cite the following reasons to consider social media in public opinion and survey research: (1) cost efficiency, (2) ease of collection of data, and (3) popularity of social media in the past few years [4].

   Twitter has been used to study public opinion on political events such as the United States Presidential Elections [26], Irish General Election [8] and the Tunisian Uprising [56]. Hu et al. [25] tried to determine positive or negative public reaction to comments made by United States Presidential candidates during debates. A comprehensive analysis would consider demographics such as gender, age, ethnicity, and location. This can enable location-based analysis such as opinion on key issues in Republican states, Democratic states and politically contested swing states.

   Location information is also key to public opinion research on social issues. The *eDrugTrends* [5] project started at Kno.e.sis research center [6], aims to process social media data to identify emerging trends in cannabis and synthetic cannabinoid use in the US. To understand the trends and public opinions on this topic, keywords such as *cannabis*, *spice*, *k2* are used to filter tweets from the Twitter stream. The location, of the corresponding Twitter users, are used to compare trends in knowledge, attitudes

[4] http://www.aapor.org/Social_Media_Task_Force_Report.htm
[5] http://wiki.knoesis.org/index.php/EDrugTrends
[6] http://knoesis.org/

and behaviours related to cannabis and synthetic cannabinoid use across US regions with different cannabis legalization policies.

2. **Personalization and Recommendation Systems:** Recently, there has been a strong interest in personalization and recommendation systems that offer recommendations for users based on their preferences and constraints. Adomavicius et al. [1] outline the importance of context, such as location and temporal information, in recommendation systems. They state that it is important to incorporate contextual information to recommend items to users under certain circumstances.

   Existing Twitter-based recommender systems do not exploit location of the user. These systems can exploit location information to provide localized recommendations. For instance, Twitter tailors *trending topics* [7] for a user based on their location if this information is provided explicitly by the user. Similarly, a News Recommender System such as [13], which utilizes the twitter stream of a user to recommend news articles, can benefit from the location of the user to recommend localized news articles to the user.

3. **Crisis Response:** As Hurricane Sandy made landfall in New York in October of 2012, people in the disaster-struck area used Twitter as their main source of communication with the rest of the world. Even government and disaster response agencies such as Federal Emergency Management Agency (FEMA) and American Red Cross used Twitter to post hurricane warnings, updates and important safety information [8].

   This rich exchange of information during a crisis by many different agencies offers an enormous opportunity to extract resource needs of people in crisis. This can, in

[7] https://support.twitter.com/articles/101125-faqs-about-trends-on-twitter
[8] https://blog.twitter.com/2012/hurricane-sandy-resources-on-twitter

turn, by used by emergency responders for effective coordination. As an illustration, consider the following tweet requesting for a resource sent during Hurricane Sandy:

*"If anybody with a portable generator can get to lower Manhattan, contact @***** *– she has a friend on a ventilator who needs your help."*

Automatic detection of resource requests and offers in conjunction with location information of Twitter users can help identify someone from *Lower Manhattan* who recently tweeted that they bought a generator. Identifying the location of a tweet at a fine granularity such as the locality (e.g. lower Manhattan) is a challenging task. While work has been done in estimating the accurate location of a tweet [16], they assume that the wide geographic region of a Twitter post is known. The geographic information of a user (and hence his posts) at a higher granularity level, such as city and state, can be obtained from our work.

While the applications of geographic information are multi fold, Twitter users may be reluctant to publish their location due to its implication on their safety or due to general privacy concerns [23]. Nevertheless, tweets are public communications that have been used extensively for research purposes. Furthermore, location information of Twitter users can be used not just for enhancing individual user experience on the web but also for the benefit of society at large. Therefore, inferencing location information of online-users is a significant task.

Location information of online users can be provided by users themselves or gleaned from their online activity. The present approaches for locating the position of Twitter users can be grouped into the following four categories:

- **Geotagging Tweets**: Twitter users can use the location service of Twitter, by en-

abling it in their web or profile settings, to add their latitude and longitude information to tweets. As shown in Figure 1.2, this approach of locating users has highest accuracy and resolution, as the precise geo-position of a user comes from his or her mobile device. On the other hand, the number of users whose location can be determined through this method, i.e. the coverage of users, is small. The reason is that very few users choose to enable this service. Recent studies have shown that less than 4% [40, 31] of tweets contain geo-spatial tags.

- **User Profile**: Twitter users can also choose to share their location information through their profile. The location field in a Twitter profile is free-form text field. While many users choose to leave it empty or enter invalid information such as *Justin Bieber's heart*, others specify location at various granularities like *city*, *state* and *country* [23]. Thus, most of the information entered in this field cannot be reverse-geocoded to the city. Cheng et al. [11] found that only 12% of users, in their dataset, shared their location at city and state level. User profile data therefore lets us estimate a user's location with reasonable accuracy, but with a small yet higher coverage compared to geotags.

- **Network based Prediction**: The lack of coverage of the previously mentioned methods has motivated research in automatic inferencing of location of Twitter users. Location of a user can be inferred using (1) network information, or (2) contents of Twitter posts. Network-based prediction algorithms use the location information of the followers and the followees in conjunction with their online interaction to predict the given user's location. As shown in Figure 1.2, the accuracy of this approach is thus lower than the accuracy of user profile as the information is not volunteered by

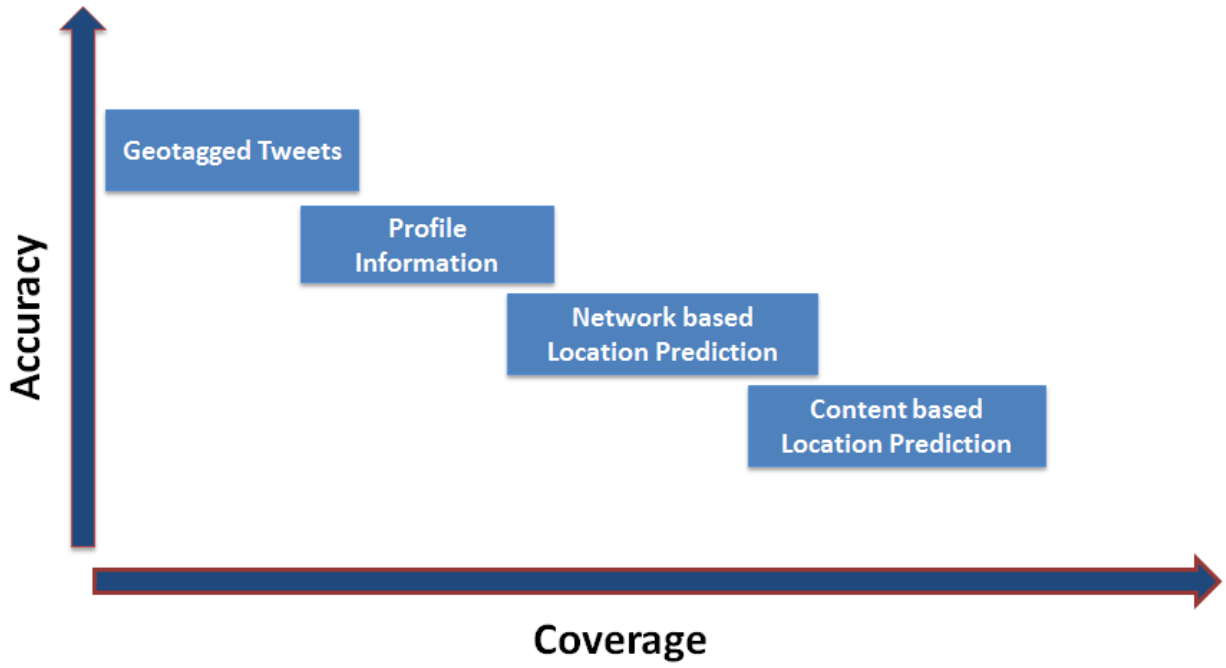

Figure 1.2: Accuracy and Coverage of Location Prediction Techniques

user but automatically gleaned from their network information.

- **Content based Prediction**: Content based approaches predict the location of user based on the analysis of the textual content of a user's tweets. Compared to the network based approaches, the content based approaches have a wider coverage as the network-based approaches can only be applied to a user who has other users in his/her network with known location whereas content based approaches can be applied to all active Twitter users.

Existing content-based approaches [10, 11] are based on the intuition that the geographic location of users influences the content of their tweets. For instance, users are likely to tweet about shops, restaurants, sports teams of their location and use location in-

dicative slang words like *howdy* (Texas). These techniques thus require substantial quantity of training examples, i.e. tweets labelled with location information. For instance, the training dataset, created by Cheng et al. [11], consists of 4 million tweets from 130,869 users whose location information is extracted from their profile. The creation of a new model, to predict the location of a different geographic region, will require the collection of a fresh set of tweets. Furthermore, these approaches fail to exploit the underlying semantics of the tweets. For example, *Boston Red Sox* and *BoSox* refer to the same team but cannot be identified as the same concept without the semantic analysis of the text. Semantic knowledge has been exploited in many Information Retrieval problems such as Text Clustering and Classification but has yet to be explored in the area of location prediction of online users. Recently, gazetteers have been used as a source of toponyms to increase prediction based on the classifiers [36]. DBpedia has been used, again as a source of toponyms, to create an unsupervised approach to predict the location of a user [30]. However, the use of an external-knowledgebase to provide location specific concepts has been missing.

This thesis addresses a major weakness of the current methods for Twitter use home location prediction by presenting a new method that exploits Wikipedia for the semantic analysis of tweets. The *home location* of a Twitter user may be considered as the location of their primary residence. While users may travel and tweet from various locations, in this work we gather evidence from their historical tweets and predict their most likely home location. The core of the method lies in extracting location-specific information from Wikipedia. It exploits the idea of local words proposed by Cheng et al. [11]. *Local Words* are words that convey a strong sense of location. For example, they found that the word *rockets* is local to Houston whereas words such as *world* and *peace* are more generic and do not exhibit an association to any particular location. Examples of local words are

Location	Local Word
Las Vegas, Nevada	casino
Grand Canyon, Arizona	canyon
San Diego, California	chargers
Yonkers, New York	yonkers
Ames, Iowa	isu
Dallas, Texas	mavs
Corpus Christie, Texas	corpus

Table 1.1: Example of Local Words

shown in Table 1.1. We extend this idea to define *Local Entities* as entities that are able to discriminate between geographic locations.

This approach relies exclusively on the tweets of users and does not require other metadata such as user's profile or network information. We build a knowledge-base for each city by using entities found in its Wikipedia page. We use the hyperlink structure of Wikipedia to weight these entities based on the degree of their association to the city. Next, we apply named entity recognition on tweets and associate the entities found with concepts in Wikipedia. Finally we use the overlap between the entities in the tweets of a user and knowledge-base of cities to predict their location.

The rest of the thesis is organized as follows. In Chapter 2, we survey the related work in the area of association of geographic information with content on the web and the use of Wikipedia, as an external knowledge-base, in Information Retrieval. In Chapter 3, we introduce Wikipedia and its hyperlink structure. In Chapter 4, we describe the architecture of our approach. In Chapter 5 we present an evaluation of our approach on a publicly available dataset and compare it to the state-of-the-art approaches. Finally, in Chapter 6, we conclude with the future work.

Related Work

Location prediction has long been a research topic across many application areas. Earlier research was pivoted around improving the results of search engines using geography as an added dimension to help rank the "usefulness" of a web page with respect to a query [15, 2, 9, 37]. In this context, the geography associated with a web page is the geographic focus of the content of the page as opposed to the physical location of the server it is stored at. In general, the geographic scope of a web page is computed either using (1) the content of the page, or (2) the geographical distribution of hyperlinks to the page.The content-based approaches associate geography with each web page by identifying location references and disambiguating them. Thus, they can only be applied to web pages that explicitly refer to one or more locations.

With the increasing number of applications based on user generated content in social media, a new problem namely location prediction of social media users has gained traction. Our research focusses on location prediction of Twitter users, based on the textual contents of their tweets, using an external knowledge-base. In this chapter, we present prior work on predicting the geographic location of a social media user. We follow it up with a discussion on earlier research that utilizes external knowledge-base in text mining.

2.1 Location Prediction of Social Media Users

For the purpose of this research, the location of an online user is considered to be his/her home location, i.e., their location of residence. Applications such as opinion analysis systems and crisis management systems generally require the residence location of a user rather than the exact location of a user (e.g their location when they are travelling). Also, there has been focus on identifying the location of an individual tweet. Generally a tweet by itself may not provide enough clues to predict its location. In this case, the home location of a user can be used as an additional feature to predict the location of a single tweet.

Similar to the location prediction of a web page, the approaches to predict the location of a user can be categorized into (1) content-based approaches that exploit user generated content such as social media posts and check-ins, and (2) network-based approaches that utilize the network information of the user, i.e, their online "friends".

2.1.1 Network-based Location Prediction of Online Users

Backstorm et al. [7] published the seminal work in predicting the location of users in a social network by exploiting the network of a user. Their dataset consists of 3.5 million Facebook users who published their address. On an average, these users had 10 friends with addresses. They used this dataset to study the relationship between geographic distance and social relationship. They found that the probability of a friendship is inversely proportional to distance. Next, they used these observations to propose a Maximum Likelihood Algorithm to predict the location of a user who had not volunteered this information in their profile. Using this algorithm, they were able to predict 67% of users' location within 25 miles of their actual location. These users had atleast 16 friends whose actual location

was known. With atleast one friend whose actual location was known, this algorithm could predict 51.38% of users within their given locations.

Rout et al. [48] formulated the geo-location prediction task as a classification task and trained an SVM classifier with features based on the information of users' followers-followees who have their location information available. For each city, the features considered were (1) number of friends in the city, (2) the size of the city, (3) city population bins, (4) triads, i.e, group of three people, and (5) reciprocated friendships.They tested their approach on a random sample of 1000 users and reported 50.08% accuracy at the city level.

McGee et al. [38] proposed FriendlyLocation, an approach based on *tie strength* between Twitter users for estimating their location. Their hypothesis was that certain relationships, such as those with a user's co-workers, contain more discriminating power than others, such as those with a news service. They analysed tweets from 1,758,101 users with known location to understand hidden patterns in the communication between users on Twitter. They used these observations to train a Decision Tree Classifier to distinguish between pairs of users likely to live close by. They reported an accuracy of 64% within 25 miles.

A major drawback of the network-based approaches is that they primarily depend on other users, in the network of a given user, whose location is known. In a social network, many users are concerned about their privacy and do not publish their actual location. Hence the content-based approaches are significant when there is no other means of identifying a given user's location.

2.1.2 Content-based Location Prediction of Online Users

Content-based location prediction approaches are grounded on the premise that the online content of a user is influenced by their geographical location. They rely on a data set of geo-tagged tweets to build a statistical model that identify words with a local geographic scope. Cheng et al. [11] proposed a probabilistic framework for estimating a Twitter user's city-level location based on the content of approximately 1000 tweets of each user. They formulated the task of identifying local words as a decision problem. They used the model of spatial variation proposed by [6] to train a Decision Tree Classifier using a hand-curated list of 19,178 words. Their approach on a test dataset of 5119 users could locate 51% of the users within 100 miles with an average error distance of 535 miles. The disadvantage of this approach was the assumption that a "term" is spatially significant to only one location/city. This challenge was addressed by Chang et al. [10] by modelling the variations as a Gaussian Mixture Model. Furthermore, their approach to identify local words did not need a labelled set of seed words. Their tests on the same dataset showed an accuracy (within 100 miles) of 49.9% with 509.3 miles of average error distance.

Eisenstein et al. [17] proposed cascading topic models to identify lexical variation across geographic locations. Using the regional distribution of words, determined from these models, they predicted the locations of twitter users. They found that slang words have a stronger regional bias as compared to standard english words. Kinsella et al. [32] addressed two problems, namely, (1) predicting the location of an individual tweet and (2) predicting the location of a user. They created language models for each location at different granularity levels of country, state, city and zipcode, by estimating a distribution of terms associated with the location.

Doran et al. [16] presented a probabilistic language model to accurately estimate the

location of a single social media post. By assuming that the "broad region" of a tweet is known, they proposed a methodology to predict the location of a tweet within few miles of its actual location. They divided a region into sub-regions and built a probabilistic language model over each sub-region. Finally, they applied geo-smoothing to improve the accuracy of their prediction. In a dataset of tweets from New York City, they could estimate the location of a tweet to within 4km with 80% probability.

Katragadda et al. [30] proposed an unsupervised approach that used gazetteer to identify location references in the contents of a user's tweets. Their approach did not require any training dataset. Yet they only focus on location entities, i.e. entities with latitude and longitude information. They used DBPedia to extract and disambiguate location references in tweets. Then they apply K-Center Clustering to predict the location of a user. A significant difference between their work and ours is that our approach is not restricted to using location reference in tweets. We consider different types of entities such as sports teams, cultural entities and transportation services.

Jalal et al. [36] used an ensemble of statistical and heuristic classifiers to predict the location of a user. These classifiers use words, hashtags and location names as features. A low level classifier, that predicts location at city-level, needs to discriminate among many locations. To alleviate that, they propose an ensemble of hierarchical classifiers that predict the location at time zone, state, region and city level. Additionally, they also train a classifier to detect travelling users and eliminate them from their test set. On a dataset of 9551 users they report an accuracy of 61%.

2.2 Role of Background Knowledge in Text Mining

Traditional text analysis algorithms can be broadly categorized into Supervised and Unsupervised Learning. Supervised Learning requires a labelled dataset used to infer a function which can be subsequently used for mapping unlabelled data. On the other hand, Unsupervised Learning techniques find hidden structure in an unlabelled dataset. These approaches are purely data-driven and do not overcome the lack of domain knowledge or common sense.

As humans, we use our general knowledge in the interpretation of any text or discourse. For instance, when we read *"Buckeye State"* in a piece of text, we use our experience to recognize it as the state of Ohio. Similarly, background knowledge can improve a machine's ability to understand and interpret text.

Background knowledge can be represented in various ways with different expressive power such as thesaurus, taxonomy and ontology. In the field of computer science, ontology is a formal representation of concepts in a domain. It contains a set of classes with each class having an associated set of properties. Additionally, an ontology relates specific concepts to more generic concepts forming an inheritance hierarchy. Enrichment of a document using ontology provides contextual information and helps in the deeper understanding of the document text. For example, ontologies such as MeSH [1], SNOWMED [2] and UMLS [3], have been extensively used in medical domain for the purpose of knowledge acquisition [45], understanding document content for secondary data analysis, relevant document retrieval [52] etc.

[1] http://www.ncbi.nlm.nih.gov/mesh
[2] http://www.ihtsdo.org/snomed-ct/
[3] http://www.nlm.nih.gov/research/umls/

WordNet [4] is an example of a lexical knowledge-base. It consists of nouns, verbs, adjectives and adverbs of English language. These words are grouped into sets of synonyms called synsets. Furthermore, each of the synsets are linked to others through conceptual relations such as hypernomy and hyponomy. Richardson et al. [46] used Wordnet to measure the conceptual similarity between words. Wang et al. [54] used WordNet synset to exploit relationship between terms that do not co-occur frequently. They showed that text clustering algorithms perform better on documents enriched with background knowledge compared to documents represented as bag-of-words.

Background knowledge has also been widely used in the task of Named Entity Recognition and Disambiguation. Gruhl et al. [21] used MusicBrainz ontology for entity spotting in informal textual content in music domain. Hassell et al.[22] use an ontology extracted from DBLP bibliography to disambiguate names of researchers appearing in a collection of DBWorld Posts.

2.2.1 Wikipedia as Background Knowledge

Wikipedia - an online collaborative encyclopedia, has been leveraged as background knowledge in many tasks. Each article in Wikipedia represents a single concept. Gabrilovich et al [19] used the content extracted from Wikipedia pages to enrich document representation for the task of text categorization. Each Wikipedia article is represented as a vector of words that appear in the article. Then, machine learning techniques are used to map text from documents to the aforementioned vector representation of Wikipedia concepts. On a test dataset consisting of documents from Reuters and OSHUMED [5], they showed

[4] http://wordnet.princeton.edu/
[5] subset of MEDLINE

that knowledge extracted from Wikipedia is very useful in categorizing short documents. Mukherjee et al. [41] proposed an unsupervised approach to perform sentiment analysis of movie reviews. They used the domain-specific information such as crew, plot and character information from the infobox of the Wikipedia page of a movie. Their system did not need any labelled data for training and achieved comparable results to the semi-supervised and unsupervised state-of-the-art systems.

Concepts in Wikipedia are organized in a category structure where each concept belongs to one category. Hu et al. [24] used the category structure of Wikipedia for the purpose of document clustering. Each word in the document is weighted using *tf-idf* and associated Wikipedia concepts and categories are retrieved. Thus a given document is represented as a vector of weighted terms occurring in the document, a vector of relevant Wikipedia concepts and a vector of categories of the concepts. Finally, partitional clustering is used to compute the similarity between the vectors of two documents. Their tests on three datasets showed that category information is useful in document clustering. The category structure of Wikipedia has been utilized by Genc et al. [20] to classify tweets. Their approach first maps each tweet to the most relevant Wikipedia concept and further leverages the category structure to find the semantic distance between the mapped concepts for classification. Kapanipathi et al. [29] used an adaptation of spreading activation theory on the category structure to determine the hierarchical interests of users based on their tweets.

Each article in Wikipedia contains links to other articles that are used to describe a concept in the main article. These links are referred to as *wikilinks* and they form a hyperlink graph. Milne et al. [55] used the link structure of Wikipedia to compute semantic relatedness between two terms using the hyperlinks found in their respective Wikipedia articles. First, they use anchor text to determine the link the Wikipedia page that maps

to a given term. Then they measure the similarity of the Wikipedia articles using Normalized Google Distance between the vector of links found in the two Wikipedia articles. Their tests showed that the accuracy of their approach did not work as well as the Explicit Semantic Analysis proposed by Gabrilovich et al [19] but required far less resources. Li et al. [35] proposed a Wikipedia based approach for Word Sense Disambiguation that did not require any labelled training dataset. They extract keyphrases and corresponding topic from Wikipedia where keyphrases are Wikipedia article titles and anchor texts of wikilinks. Next, they extract phrases from the input document and map unambiguous keyphrases to those extracted from Wikipedia. These phrases provide context to the disambiguator that computes similarity between Wikipedia articles to choose the most appropriate sense of the word.

Knowledge-base Creation

As explained in Section 2.2, background knowledge has been incorporated in many Natural Language Processing and Information Retrieval tasks. Wikipedia is a rich source of information for geographic locations and can be exploited as background knowledge for the task of location prediction. In this chapter, we discuss our approach to extract location-specific knowledge from Wikipedia. For the creation of the knowledge-base, our goal is to (1) identify all the entities that can be used to describe a given location, and (2) compute the degree of their association to the location. The presence of local entities in a user's tweets, their frequency and localness measure are used to rank the top-k locations of a user.

3.1 Wikipedia

> *"Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we are doing."*
> - Jimmy Wales (Founder of Wikipedia)

In 2001, Jimmy Wales and Larry Sanger founded Wikipedia using the concept of Wiki pioneered by Ward Cunningham in 1995. *Wiki* is a web application that allows people to create or update content in collaboration with others. Wikipedia, a publicly available online

Figure 3.1: Internal links of Wikipedia

collaborative encyclopedia, is the most popular wiki on the web. It has been a prominent source of knowledge for humans as well as machines. As of July 2014, Wikipedia is available in 287 languages. It has 18 billion page views and approximately 500 million unique visitors each month. The English Wikipedia comprises of approximately 4.6 million articles. It is open-access and updated regularly by active contributors. Wikipedia articles are comprehensive and well-formed with each article describing a single topic or entity [24]. In comparison to knowledge-bases such as MeSH and Wordnet, Wikipedia is not domain specific and provides broad coverage of topics. Wikipedia is open to all members of the public. Anybody can contribute to an article, correct errors and/or compensate for any skewed views in an article. Eric Raymond claimed, in the context of software development, that "given enough eyeballs, all bugs are shallow". This is also applicable to Wikipedia.

To describe an entity in an article, it can be linked to the Wikipedia page of the said entity. For example, consider the snippet from the Wikipedia page of "Semantic Web", shown in Figure 3.1. A link to the Wikipedia page of *Resource Description Framework*

is added to describe that topic. These links are referred to as *wikilinks* or *internal links* [1]. The aim of these links is to enhance a user's understanding about the topic of the page. Wikipedia has established some guidelines on adding internal links to a Wikipedia page.

As per Wikipedia policy:

> *"In general, links should be created to relevant connections to the subject of another article that will help readers understand the article more fully. This can include people, events, and topics that already have an article or that clearly deserve one, so long as the link is relevant to the article in question."*

The collaborative nature of Wikipedia ensures that the guidelines outlined above are followed by the authors. Consequently, there is less redundant information and all relevant links are present. Apart from allowing a user to navigate to the pages of related topics, the internal links also creates a hyperlink structure, that allows machines to use Wikipedia as a knowledge base of semantically linked entities with the underlying assumption that an article's main subject is soundly and centrally related to the linked articles' main subjects [42]. For the task of location prediction of Twitter user, we use Wikipedia as the knowledge-base for the following reasons:

- The hyperlink structure of Wikipedia provides links for each page that are topically relevant to that page. In other words, the internal links of a Wikipedia page are semantically related to it (or a portion of it) [28, 42]. Thus, we hypothesize that for each location, the internal links in its Wikipedia page represent entities that are relevant to it. For example, a link from the Wikipedia page of *San Francisco* to *Golden Gate Bridge* implies that the latter is relevant to the former.

[1] http://en.wikipedia.org/wiki/Help:Link#Wikilinks

- Named Entity Recognition tools such as DBpedia Spotlight and Zemanta, map the entities annotated in a piece of text to their corresponding Wikipedia entities. This allows us to map the entities found in a user's tweets to the entities in our knowledge-base extracted from Wikipedia.

- We use Wikipedia instead of a gazetteer because of the broad coverage of categories in a Wikipedia page. While a Wikipedia article on a location may contain several sections like *History*, *Geography*, *Cityscape*, *Neighbourhoods*, *Climate*, *Sports* and *Culture*, gazetteers are limited to the geographic features of a region.

- We find that Wikipedia contains dedicated pages for all cities in United States with population greater than 5000. Also, Wikipedia contains pages for locations at different granularity levels such as state, county and locality which will allow us to extend this work to predict the location of a user at different granularity levels.

Formally, the **Wikipedia Hyperlink Graph** can be represented as a directed graph $G = (V, E)$ with a set of vertices V that represents a subset of all the Wikipedia pages and a set of edges E, where $E \subseteq V \times V$. There is a directed edge (v_1, v_2), if there is a link from Wikipedia page v_1 to v_2. For a given vertex v_i, $O(v_i)$ is the set of entities mentioned in the Wikipedia page v_i, i.e, $O(v_i)$ are the vertices that have an edge from v_i.

3.2 Local Entities

Local Entities are entities that can discriminate between geographic locations. Thus, intuitively, *Statue of Liberty* may be considered a local entity with respect to New York City on account of it being a famous landmark in New York City whereas *iPhone*, a smart phone

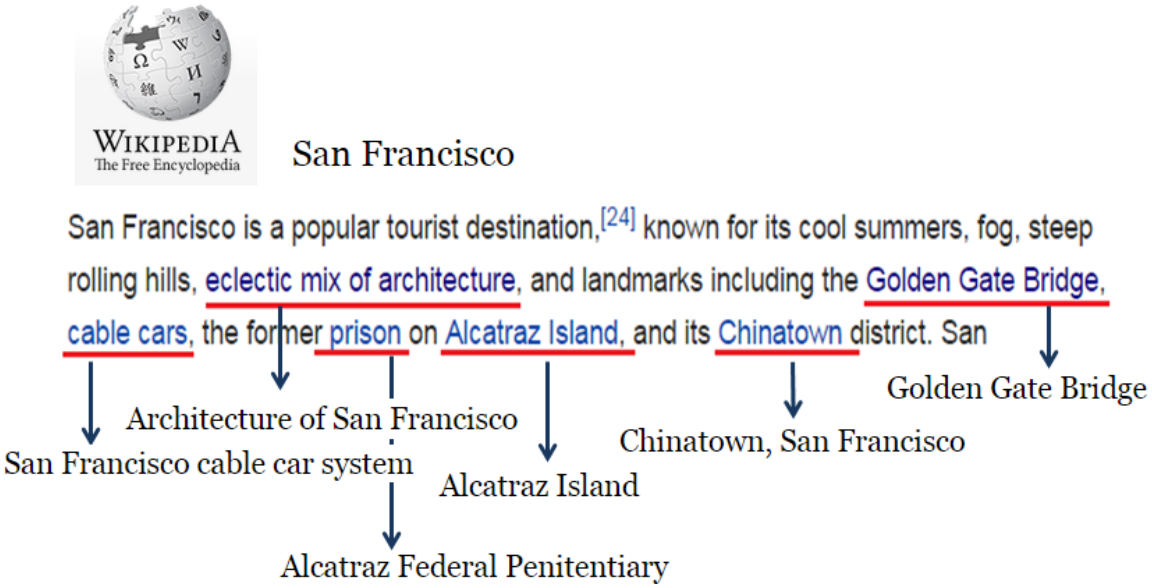

Figure 3.2: Local entities of San Francisco

with over 63.2 million users across United States, may not be considered as a local entity.

The internal links of a Wikipedia page represent entities that are topically relevant to the main page and are established by using the collective wisdom of Wikipedia contributors [42, 35]. Thus we consider the entities mentioned in a Wikipedia page of a city c as the local entities of c. From the hyperlink structure, the local entities are the outgoing links $O(c)$ from each Wikipedia page of city c. Furthermore, the local entities vary in the degree of their localness with respect to the location. For example, the Wikipedia page of *San Francisco* contains links to *San Francisco Bay Area* and *United States*. Also, an entity may not be local to just one city. For instance, *San Francisco Bay Area* is also found in the Wikipedia pages of *Alameda, California*, *Los Altos, California*, *Sacramento, California* etc. Figure 3.2 shows local entities from a snippet of Wikipedia page of San Francisco.

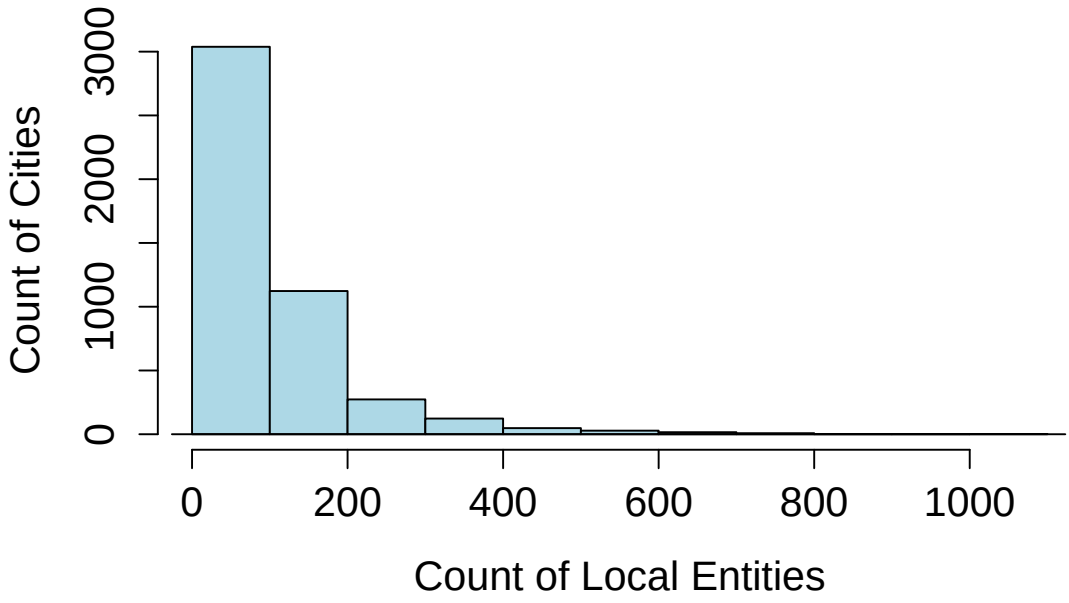

Figure 3.3: Count of Local Entities for cities in US with population > 5000

3.3 Localness Measure of Entities

The number of local entities of a city may vary depending on the size of the city and the information available in Wikipedia. For instance, *San Francisco, California* has 717 local entities whereas *Fairborn, Ohio* has 110 local entities. Figure 3.3 shows the count of local entities across the cities of United States with population > 5000, as published in the census estimates of 2012. Each of these local entities is not equally local to the city. As an example, consider the following snippet from the Wikipedia page of *San Francisco, California*:

The San Francisco 49ers of the National Football League(NFL) were the longest-tenured professional sports franchise in the city.

The San Francisco 49ers are the American football team located in San Francisco whereas the National Football League is one of the professional sports leagues in North America. Clearly, the entity *San Francisco 49ers* has a higher degree of localness with respect to San Francisco than the entity *National Football League*. Similarly, each city has local entities with high relatedness to the city such as the sports team based out of the city, they mayor of the city and famous landmarks of the city. On the other hand, every city also has entities that have less or no relatedness to the city. In other words, these entities do not help in discriminating between cities. To this end, our goal is to score each local entity with respect to a city such that the score reflects the discriminating power of the entity.

We experiment with four measures, that exploit the hyperlink structure of Wikipedia, to score the localness of an entity with respect to a city. These measures can be classified into three categories namely (1) association measure, (2) graph-based measure, and (3) semantic overlap measure. Association measures have been commonly used in computing the relatedness of two words based on their occurrences in a large corpus such as the web [55, 27]. We use the same idea to measure the relatedness between a city and a local entity based on their occurrences in the entire Wikipedia dump. Next, we explore graph-based measures as the Wikipedia Hyperlink Graph allows us to conveniently represent a city and its local entities as a graph. Betweenness centrality of a node in a graph has been used in social networks to compute the importance of an actor in a network [33]. We use this measure to compute the relative importance of a node representing a local entity in the graph of a city. Finally, we investigate semantic overlap measures that are based on the idea that

higher the overlap between concepts found in the Wikipedia pages of a city and a local entity, higher is the degree of localness of the entity.

3.3.1 Association Measure

An *association* between two words indicates statistical dependence between them. There are many association measures that have been used in NLP tasks such as collocation extraction and multi-word expression extraction [44]. The intuitive basis for using an association measure, to establish localness of an entity, is that higher is the localness of an entity with respect to a city higher will be its association to the city. Pointwise Mutual Information (PMI) [12] is a standard measure of association between two events. Given two random variables, PMI is a measure of how much the actual probability of their occurrence differs from what is expected based on the probabilities of their individual occurrences assuming independence. For measuring the PMI of a local entity and a city, we consider the entire Wikipedia dump as our corpus. We define the PMI of a city and its local entity as:

$$\begin{aligned} PMI(c,e) &= \log_2 \frac{P(c,e)}{P(c)P(e)} \\ &= \log_2 \frac{\frac{|C \cap E|}{|W|}}{\frac{|C|}{|W|}\frac{|E|}{|W|}} \end{aligned} \tag{3.1}$$

where c is the city, e is a local entity of the city, $|C|$ is the number of Wikipedia articles that contain the city c, $|E|$ is the number of Wikipedia articles that contain the entity e, $|C \cap E|$ is the number of Wikipedia articles that contain both the entity and the city and finally $|W|$ is the entire Wikipedia dump.

In this context, the occurrence of an entity in a Wikipedia page refers to the presence of a wikilink to the entity from the said page. The individual probabilities of the city $P(c)$ and the local entity $P(e)$ are computed as the fraction of all the Wikipedia articles that contain the city and the local entity respectively. The joint probability $P(c, e)$ is the fraction of all the Wikipedia articles that contain both the city and the entity.

3.3.2 Graph-based Measure

Centrality measures have been used as indicators to identify the most important vertices within a graph. Centrality measures based on the degree of a node and the shortest paths between nodes are commonly used to determine the relative importance of a node in a graph. Betweenness Centrality (BC) [18] of a node measures the prominence of the node relative to the rest of the nodes in the network. A high betweenness centrality score of a vertex in a graph indicates that it lies on considerable fraction of shortest path connecting others.

The graph of local entities for each city is pruned from the Wikipedia hyperlink graph and consists of the entities present in the corresponding city's Wikipedia page ($O(c)$). Formally, the graph for a city c is represented as $G_c = (V_c, E_c)$ where vertices $V_c \in (c \cup O(c))$ and edges $E_c \in V_c \times V_c$. There is an edge from v_{c_i} to v_{c_j} if the Wikipedia page of v_{c_i} has a link to entity v_{c_j}. An example of a subgraph is as shown in Figure 3.4. The nodes in this graph are the entities mentioned in the Wikipedia page of *San Francisco*. We draw edges between entities based on the entity occurrences in their respective pages. For instance, an edge between *Golden Gate* and *Bay Area* is indicative of the presence of the latter in the former's Wikipedia page.

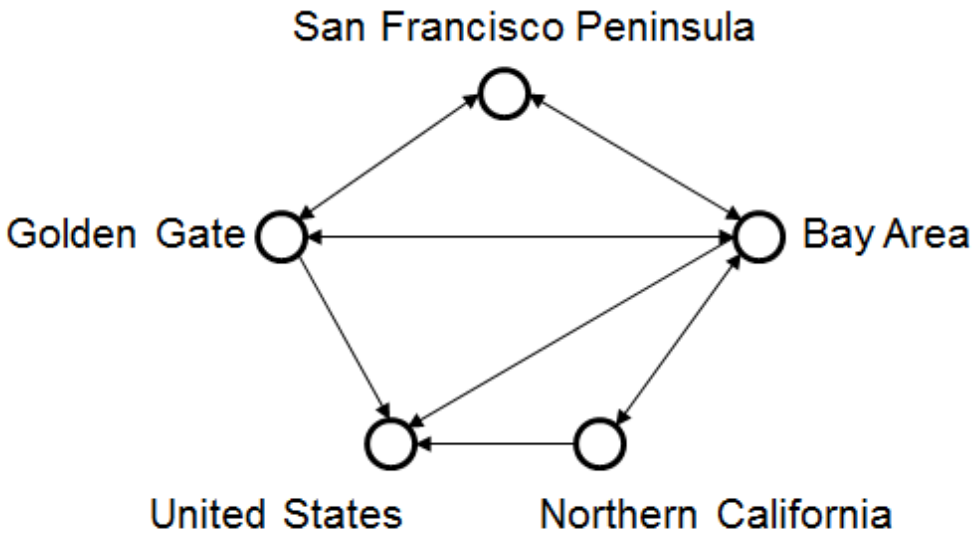

Figure 3.4: Pruned Subgraph of San Francisco

Betweenness Centrality is defined as follows:

$$C_B(c, e) = \sum_{e_i \neq e \neq e_j} \frac{\sigma_{e_i e_j}(e)}{\sigma_{e_i e_j}} \tag{3.2}$$

where $e_i, e, e_j \in O(c)$, $\sigma_{e_i e_j}$ represents the total number of shortest paths from e_i to e_j and $\sigma_{e_i e_j}(e)$ is the number of shortest paths from e_i to e_j through e. Furthermore, we normalize the measure by dividing C_B by $(n-1)(n-2)$ which is the number of pairs of nodes not including e with n being the number of nodes in the directed graph. Thus, betweenness centrality of each node is a number between 0 and 1.

3.3.3 Semantic Overlap Measure

SemRank [4], a search results ranking system, measures the relatedness between concepts with the intuition that related concepts are connected to similar entities. Similarly, we use the Wikipedia hyperlink graph to determine the extent of relatedness between a city and an entity. We term this as *Semantic Overlap*. We use two standard set based measures to compute the semantic overlap between a city and an entity, namely, the Jaccard Index and the Tversky Index.

Jaccard Index measures the overlap between two sets and is normalized for their sizes. We use this measure to find the similarity between a city and its entities. For example, to compute the localness of *Golden Gate Bridge* to *San Francisco*, we compute the Jaccard Index of the two sets containing the entities from the Wikipedia page of *Golden Gate Bridge* [2] and *San Francisco* respectively. Jaccard Index for a city c and entity e ($e \in O(c)$) is defined as shown in Equation 3.3.

$$jaccard(c, e) = \frac{|O(c) \cap O(e)|}{|O(c) \cup O(e)|} \tag{3.3}$$

The idea behind using Jaccard Index is that larger the overlap between the entities associated with a specific entity and entities associated with a city, higher is the localness of the entity with respect to the city. The range of Jaccard Index is between 0 and 1.

Tversky Index is an asymmetric similarity measure between two sets [53]. While the Jaccard Index determines the similarity between a city and a local entity, a local entity generally represents a part of the city. For example, consider the local entity *Boston Red Sox* of the city *Boston*. *Boston Red Sox* is the baseball team of Boston and will not completely

[2] Entities of Golden Gate Bridge are the wikilinks appearing in the Wikipedia page of Golden Gate Bridge

overlap with all the entities of *Boston* which are from different categories like *Climate*, *Geography* and *History*. Thus we use Tversky Index which is a unidirectional measure of similarity of the local entity to the index.

The Tversky Index is defined as shown in Equation 3.4.

$$ti(c, e) = \frac{|O(c) \cap O(e)|}{|O(c) \cap O(e)| + \alpha|O(c) - O(e)| + \beta|O(e) - O(c)|} \tag{3.4}$$

where we choose $\alpha = 0$ and $\beta = 1$, with no weight given to the entities of the city. Thus for every entity in its page that is not found in the Wikipedia page of the city c, we penalize the local entity e. The range of Tversky index is between 0 and 1.

Knowledge-base Enabled Location Prediction

Previous research that address the problem of location prediction, have established that the content of a user's posts reflects his/her location. They rely on *words* that may be deemed to have a geospatial dimension. Use of these words with high frequency is indicative of a user's location. In our approach, we move a step further by considering *entities* in the tweets of a user. Our hypothesis is that users are likely to tweet about *entities* that are local to them. For instance, people are inclined to tweet about, among other things, landmarks, restaurants, shopping malls, politicians, sports teams etc in their neighbourhood giving away their location. Some sample tweets are shown in Table 4.1.

In Chapter 5, we discussed the creation of a knowledge-base of cities consisting of their local entities. Furthermore, these local entities are scored based on the degree of their association with the city. In this chapter, we create a semantic profile of each user consisting of Wikipedia entities found in their tweets. Subsequently, we use a subset of these entities, which are also present in our knowledge-base, along with their localness scores and frequency in the user profile, to rank the *top-k* locations of a user.

Tweet	Local Entity	Description
"Just drove around Golden Gate Park two times trying to get in."	Golden Gate Park	Urban park in San Francisco
"Just told the boys that we'll be going to see the Red Sox on Sunday."	Boston Red Sox	Baseball team based in Boston
"And Nate Silver is on air on @WNYC! Woot."	WNYC	Public radio station located in New York City
"Waiting for BART from SFO to Powell St. Exit. Should be 30 minutes"	Bay Area Rapid Transit	Rapid Transit System serving bay area

Table 4.1: Tweets containing local entities

4.1 Building User Profile

With the rise of social media, Twitter has become an important medium of communication. Twitter users can broadcast their thoughts to the world using short textual updates called *Tweets*. Tweets may also contain links to images or videos. Users can re-post a tweet which is referred to as a *retweet*. Users can use hashtags (e.g. #WSU) to indicate the main topic of their tweet. The length of each tweet is limited to 140 characters. The length limitation of a tweet has resulted in the use of informal language and slang terms in tweets. Twitter content does not follow English grammar and contains unconventional spellings and acronyms. Additionally, unlike scientific documents or newspaper articles, tweets also contain sarcasm, irony and ambiguity.

People use Twitter to post about their daily mundane activities that they would otherwise not share. These activities range from a visit to a local park to riding a metro bus.

People also share their interests, likes and dislikes. Our goal is to build a semantic profile of a Twitter user that captures such activities and interests present in a set of their historic tweets. Building this profile is a two-step process of (1) entity recognition and (2) entity scoring.

Entity Recognition is the process of recognizing information like people, organization, location, and numeric expressions from natural language text. It is an important sub-task of Information Extraction. The genre of text, such as scientific, informal and journalistic, impacts the precision and recall of the entity recognition algorithm. Earlier approaches relied on hand-crafted rules. Later, supervised and unsupervised learning techniques were used for automatic recognition of entities. These learning techniques use word level features, patterns and gazetteers for the recognition and classification.

As explained in [47], the length of tweets ($<=$ 140 characters) and the informal nature of their content make the task of entity recognition on tweets non-trivial. Recently, a lot of research has focussed on automatic identification of entities (or concepts) in a tweet and linking them to their corresponding Wikipedia articles. In this work, our focus is on the location prediction of Twitter users. Hence, we use the APIs available for Entity Recognition. In [14] authors have compared three different state of the art systems, namely, Dbpedia Spotlight [39], Zemanta[1] and TextRazor[2] for entity recognition in tweets. These results are summarized in Table 4.2.

We used Zemanta for the following reasons:

- It has been shown to be superior to others.
- Zemanta's web service[3] also links entities from the tweets to their Wikipedia articles.

[1] http://developer.zemanta.com/
[2] http://www.textrazor.com/technology
[3] http://developer.zemanta.com/docs/suggest/

Extractors	Precision	Recall	F-Measure	Rate Limit
Spotlight	20.1	47.5	28.3	N/A
TextRazor	**64.6**	26.9	38.0	500/day
Zemanta	57.7	31.8	**41.0**	**10,000/day**

Table 4.2: Evaluation of Web Services for Entity Resolution and Linking

This allows an easy mapping between the Zemanta annotations and our knowledge base extracted from Wikipedia.

- The web service provides co-reference resolution for the entities.
- Zemanta provides a higher access rate limit of their API calls to 10,000 per day for research purposes.[4]

Entity Scoring entails scoring each local entity in a user's tweet using the frequency of their occurrence in the tweets and their localness measure in our knowledge-base. Formally, we define the profile of a user u as

$$P_u = \{(e, s)|e \in W, s \in \mathbb{R}\} \tag{4.1}$$

where W denotes the set of all Wikipedia entities and s is the frequency of mentions of entity e by user u. Frequency of an entity indicates the significance of the entity to the user.

[4]We thank Zemanta for their support

Local Entity extracted from Tweets	Localness Measure - Tversky Index	Frequency of Local Entity in Tweets
Las Vegas Boulevard	0.32727	11
Las Vegas	1	3
Fremont Hotel and Casino	0.26315	3
Nevada	0.09945	6
Golden Nugget Las Vegas	0.16279	2
Las Vegas Valley	0.23764	1
Pahrump, Nevada	0.10714	2
McCarran International Airport	0.03642	5
Binion's Gambling Hall and Hotel	0.10204	1
Spring Mountains	0.08	1
United States	0.01996	2

Table 4.3: Example of $locScore$ of a user with respect to the city Las Vegas

4.2 Location Prediction

We compute an aggregate score based on all the local entities found in the user profile. In other words, to estimate the location for a user u with profile P_u, for each location c with knowledge base K_c, we find the intersection of the set of entities I_{cu} associated with the user profile and the local entities of the city c. Next, we use the following equation to estimate the score for each city for a user.

$$locScore(c, u) = \sum_{j=1}^{|I_{cu}|} locl(c, e_j) \times s_{e_j} \tag{4.2}$$

where $e_j \in I_{cu}$, $locl(c, e_j)$ is the localness score of the entity e_j with respect to the city c, determined by one of the localness measure explained in Section 3.3. s_{e_j} is the score of the entity in the user profile P_u. The city for the user is determined by ranking the cities based on the $locScore(c, u)$ in descending order. An example of $locScore$ for a user with respect to the city Las Vegas is shown in Table 4.3.

Implementation and Evaluation

As shown in Figure 5.1, our approach comprises of three primary components:

- *Knowledge Base Generator* extracts local entities for each city from Wikipedia and scores them based on their relevance to the city.
- *User Profile Generator* generates a semantic profile of a user from a set of their historic tweets.
- *Location Predictor* uses the output of User Profile Generator and Knowledge Base Generator to predict the location of the user.

5.1 Implementation

To create our knowledge base, we consider all the cities of United States with population greater than 5000, as published in the census estimates of 2012. From the census estimates, we only include the locations listed as *city* and ignore the locations labelled as *village*, *town*, *county* or *CDP(Census Designated Place)*.

The entire collection of Wikipedia articles is available as an XML dump[1]. The Wikipedia

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

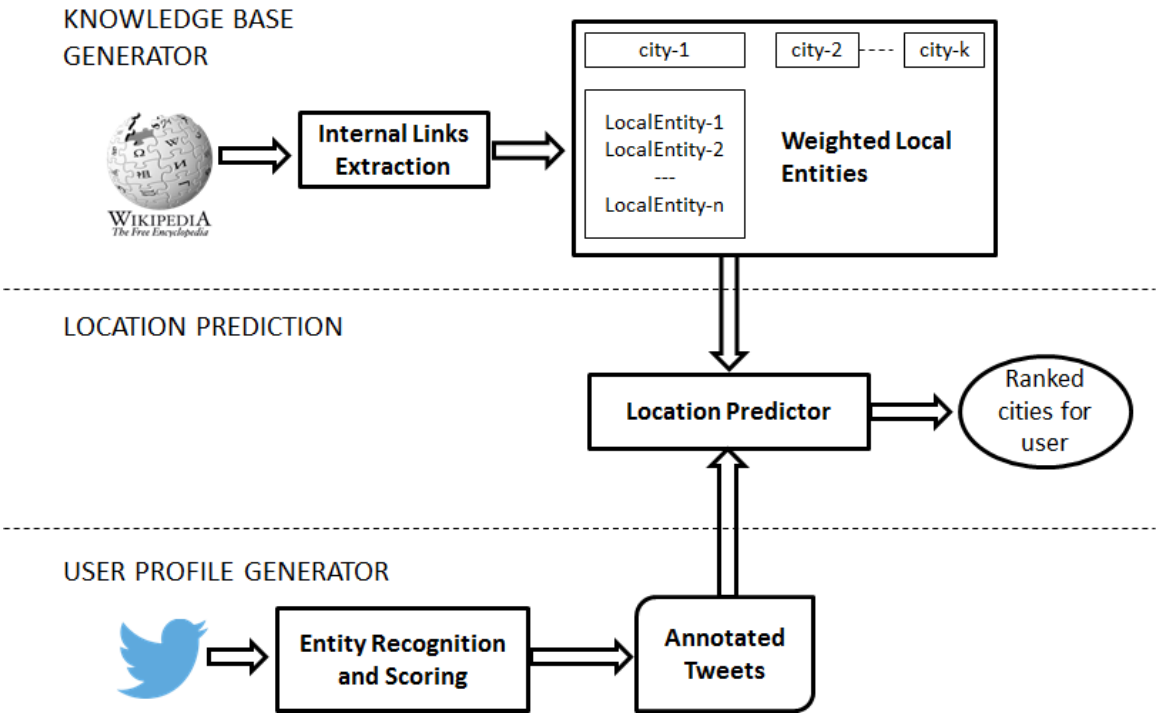

Figure 5.1: Location Prediction using Wikipedia

pages of two cities *Irondale, Alabama* and *Mills River, North Carolina* are marked as stubs[2] and hence are not included in our knowledge base. Although a Wikipedia page does not link to itself, we include the name of each city in its knowledge base. Finally, we have a knowledge base with 4,661 cities and 500,714 entities. To compute the distance between the actual and the predicted location we extract the latitude and longitude information of each city in our knowledge base from the infobox[3] of their corresponding Wikipedia page. We used Bliki engine - a java-based Wikipedia API [4] to parse the Wikipedia dump. It is a parser library for converting Wikipedia wikitext notation to HTML. It provides helper classes that can be used to extract the internal links of a given Wikipedia page. The convention to name a Wikipedia page of a city is *City Name, State Name*. For instance the Wikipedia page of the city of Fairborn in Ohio is *Fairborn, Ohio*. Pages of some cities such as *Houston* and *San Francisco* do not follow this convention and are named after the city alone. In such cases *City Name, State Name* are redirected to *City Name*. For example, *Houston, Texas* redirects to *Houston* in Wikipedia.

5.2 Evaluation

First, we compare the four localness measures explained in Section 3.3 and then use the best performing measure to evaluate against the state-of-the-art content based approach for location prediction.

[2] http://en.wikipedia.org/wiki/Wikipedia:Stub
[3] http://en.wikipedia.org/wiki/Help:Infobox
[4] http://code.google.com/p/gwtwiki/

5.2.1 Dataset

For a fair comparison of our approach against the state of art approaches, we use the dataset published by Cheng et al [11]. The dataset was collected from September 2009 to January 2010 by crawling through Twitter's public timeline APIs[5]. The dataset contains 5119 active users, from the continental United States, with approximately 1000 tweets of each user. The users' locations are listed in the form of latitude and longitude coordinates which is generally more reliable than the profile information. Spammers and bots are filtered to ensure a clean dataset. Additionally, we remove the word "RT" (referring to a re-tweet) from the tweets. We do this because Zemanta annotated "RT" in re-tweets incorrectly as *RT (TV Network)*[6] which affected the results as it is one of the local entities in our knowledge base.

5.2.2 Evaluation Metrics

We use Average Error Distance and Accuracy, as defined by Cheng et al. [11], as the two metrics to evaluate our approach. Error distance is the distance(in miles) between the actual location of the user and the estimated location by our algorithm. The Error distance for a user *u* is defined as:

$$ErrorDist(u) = distance(loc_{act}(u), loc_{est}(u)) \tag{5.1}$$

[5] http://search.twitter.com/
[6] http://en.wikipedia.org/wiki/RT_(TV_network)

where $loc_{act}(u)$ is the actual location of the user, $loc_{est}(u)$ is the location predicted using our algorithm and the $distance$ function computes the straight line distance [7] between the two locations.

Average Error Distance (AED) is the average of the error distance across all users. The AED across a set of users U is defined as:

$$AED(U) = \frac{\sum_{u \in U} ErrorDist(u)}{|U|} \tag{5.2}$$

Accuracy (ACC) is the percentage of users identified within 100 miles of their actual location. It is defined as:

$$ACC(U) = \frac{|\{u | u \in U \wedge ErrorDistance(u) \leq 100\}|}{|U|} \tag{5.3}$$

5.2.3 Baseline

We implement a baseline system which considers all the entities of a city to be equally local to the city. To predict the location of a user, we compute the score for each city by aggregating the count of local entities of the city found in the user's tweets and selecting the city with the maximum score.

[7] http://en.wikipedia.org/wiki/As_the_crow_flies

Method	ACC	AvgErrDist (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.408	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
JC	53.21	433.62	67.41	73.56	78.84
TI	**54.48**	**429.00**	68.72	74.68	79.99

Table 5.1: Location Prediction using Local Entities

5.2.4 Results

Table 5.1 reports the Accuracy and the Average Error Distance for location prediction using the (1) Baseline, (2) Pointwise Mutual Information (PMI), (3) Betweenness Centrality (BC), (4) Semantic Overlap Measures - Jaccard Index (JC), and (5) Semantic Overlap Measures - Tversky Index (TI) . We see that Tversky Index is the best performing localness measure with approximately 55% accuracy and 429 miles of AED. The accuracy is doubled compared to the baseline approach. However, compared to Jaccard Index, there is only a slight improvement in accuracy from 53.21% to 54.48% and decrease in AED from 433 to 429 miles. Furthermore, 27% of the users were located exactly at the city level.

By ranking the cities for each user, based on the descending order of localness scores, we have also evaluated the accuracy of the approach at *top-k* ranks. Similar to accuracy, *accuracy@top-k* is calculated by the number of users whose home locations are determined correctly within the *top-k* locations in the generated ranked list of locations for the user, within an error distance of 100 miles. The AED @top-k is computed using distance between the closest predicted location@*top-k* to the actual location of the user. Figure 5.2 shows the change in accuracy across *top-8* locations determined using Tversky Index.

In order to calculate the error distance for a particular user for *top-k*, we picked the closest possible location predicted by our approach to the original location of the user

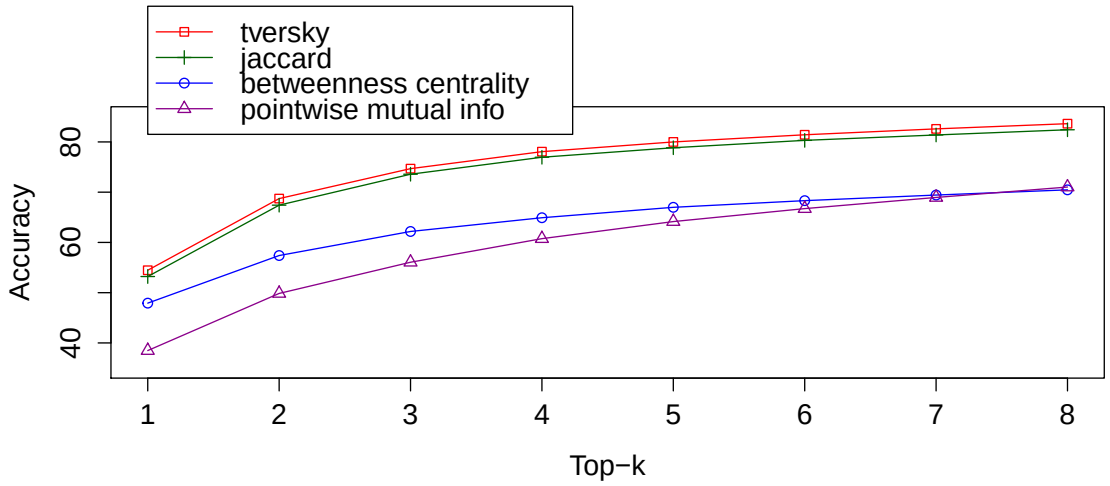

Figure 5.2: Top-k Accuracy

	City	**ErrorDistance<=50**	**ErrorDistance<=100**
Accuracy	54.65%	60.63%	63.44%

Table 5.2: Location prediction results of top 100 cities

within the *top-k* results. The error distance to this closest location is calculated and averaged across all the users to result in AED@top-k. Figure 5.3 shows that the AED decreases with inclusion of more top locations and similar to *accuracy@top-k*, Tversky Index performs the best.

We applied our algorithm for users in the top 100 most populated cities of United States. In the test dataset, there were 2172 users from these cities. Table 5.2 shows the accuracy of the algorithm for predicting the location of users from these cities, using local entities ranked using Tversky's Index. We see that our approach could predict 54% of the users at exactly the city level and 60% of the users could be located within 100 miles of their actual location.

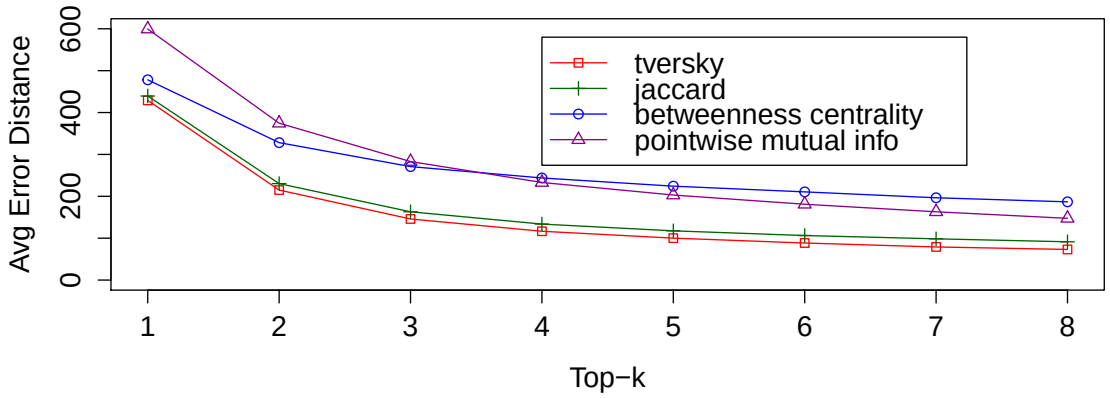

Figure 5.3: Average Error Distance

Method	ACC	AvgErrDist (in Miles)
Cheng et al.[11]	51.00	535.564
Chang et al.[10]	49.9	509.3
TI	**54.48**	**429.00**

Table 5.3: Location prediction results compared to existing approaches

5.2.5 Comparison with Existing Approaches

For the location prediction task based on user's tweets, the state of the art approaches are purely data-driven. We have evaluated our approach on the same dataset as Cheng et al. [11] and Chang et al. [10]. As reported in Table 5.3, our approach performs better in terms of both the accuracy and the average error distance. Also, note that the other approaches are based on a training dataset of 4.1 million tweets while our approach is based exclusively on Wikipedia.

5.3 Discussion

In this section, we discuss the effect of number of local entities, in the user's tweets, on the accuracy of location prediction. Furthermore, we discuss with examples the pitfalls of each localness measure and the intuition behind the best performing measure.

5.3.1 Impact of annotated entities

Figure 5.4 shows the count of all entities in the dataset annotated by Zemanta and Figure 5.5 shows the count of distinct local entities found in the tweets of users to predict their location. Note that these figures represent the predictions made using Tversky Index. From Figure 5.5, we see that when the number of local entities mentioned in the tweets are less than 5, the prediction drops by more than 12% (48% accuracy) compared to the overall accuracy of predictions. On the other hand, a prediction made on the basis of higher number of local entities is more reliable. The predictions made on the basis of 10 or more local entities were able to locate 66% of the users within 100 miles and 51% of the users within 20 miles.

5.3.2 Performance of Localness measures

We predict the location of a user based on the count of occurrences of local entities in their tweets and the localness measure of the entities with respect to a city. The pointwise mutual information measure of a city and its local entity is not normalized, making it sensitive to the count of their occurrences in the Wikipedia corpus. Consequently the absolute PMI scores of the local entities of a city like *Glen Rock, New Jersey* is higher than those of *San Francisco* because of the low occurrence of former as compared to the latter, in the

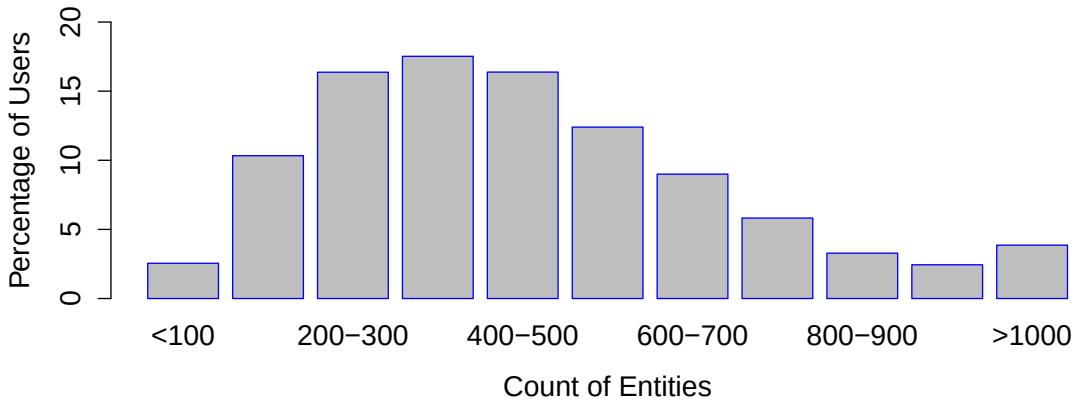

Figure 5.4: Percentage of users with the count of Wikipedia Entities extracted from their tweets

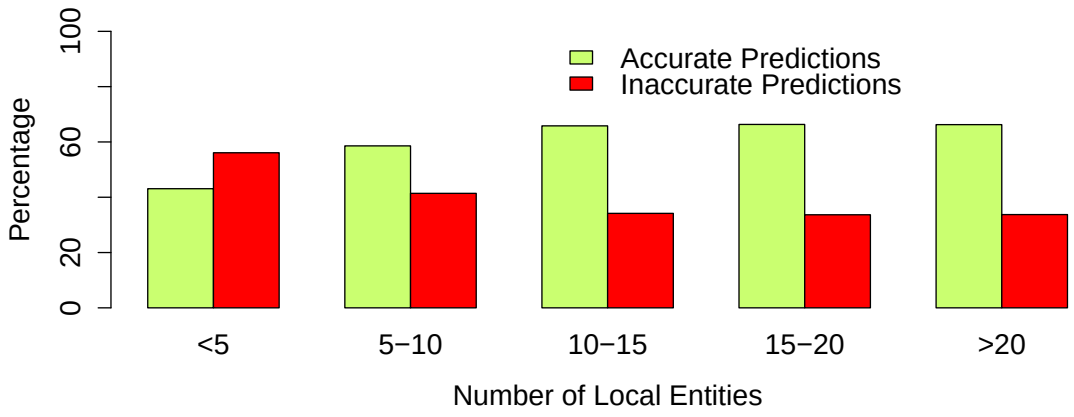

Figure 5.5: Predictions based on the number of Local Entities in users' tweets

Wikipedia corpus. This results in the location prediction to be skewed towards the cities that occur less frequently in the Wikipedia articles. Nevertheless, the prediction results using PMI show a significant improvement over the baseline. The localness of entities computed using betweenness centrality and the semantic overlap measures are normalized and yield better results than PMI.

The betweenness centrality of a node is based on the number of times a node occurs in the shortest path between two other nodes. We find that some entities which may not be local, get ranked higher because there are multiple shortest paths through them. Consider the snippet from the Wikipedia page of *Livingston, New York*, shown below:

> *The residents of Livingston are descended from people of many nations, including:*
>
> - *People from Oklahoma and other parts of the United States of America.*
> - *Hindus, Sikhs and Muslims from India and Pakistan. Livingston has one of the largest communities of Sikhs in the United States.*
> - *Mennonites from Germany and Russia.*
> - *Armenians from Middle East*

The underlined entities contain the shortest path to the rest of the entities of the city through *United States* thus increasing the importance of *United States* in the graph. Consider another example of the city *Endicott, New York*[8]. A section of the Wikepedia page of this city describes *IBM* and related entities like *Punched card* and *Circuit Board*. When we build a graph of the city, the shortest path between the *IBM* related entities and the rest of

[8] http://en.wikipedia.org/wiki/Endicott,_New_York

the entities of the city, is through *IBM*. This increases its betweenness centrality measure compared to the rest of the other local entities. As a result, when entities like *United States* or *IBM* occur frequently in a user's tweets, they lead to incorrect location prediction.

The idea behind using Jaccard Index is that larger the semantic overlap between the Wikipedia page of a city and an entity, higher is the localness of that entity with respect to the city. Thus it overcomes the disadvantage of Betweenness Centrality and is successful in assigning less localness to the more general entities like *IBM* and *United States*. However, we observe that it under-performs in measuring the localness of entities with fewer number of entities in comparison to the city. For example, consider the two entities *Eureka Valley, San Francisco* and *California*. Both are local entities of the city *San Francisco*. Intuitively, we would expect *Eureka Valley, San Francisco* (a residential neighbourhood in San Francisco) to be more local than *California* with respect to the city *San Francisco* but with Jaccard Index the result is opposite. Note that *San Francisco* has 717 entities, *Eureka Valley, San Francisco* has 36 entities and *California* has 940 entities.

The aforementioned problem is countered using the Tversky Index where the localness measure of an entity is highest when all of its entities are also present in the city. Furthermore, the localness of a local entity only diminishes for entities in its page not present in the city. Therefore, in the above example it is able to assign a higher degree of localness to *Eureka Valley, San Francisco* than *California* with respect to the city *San Francisco*. This approach to ranking the entities performs better than Jaccard's index with improved accuracy and lower average error distance. Table 5.4 shows examples of local entities from the tweets of users in the dataset used to predict their location. Figure 5.6 shows the local entities of San Francisco scored using Tversky's index.

City	**Entities**
New York City	New York City, Brooklyn, Harlem, Queens, New York Knicks, The Bronx, Manhattan, National Football League, American Broadcasting Company, Train station, Rapping, Times Square, Fox Broadcasting Company, Broadway theatre, New York Yankees, Staten Island, Brooklyn Nets, Amtrak, Hudson River, Macy's Thanksgiving Day Parade
Houston	Houston; Houston Texans; NASA; Houston Astros; Interstate 45; Houston Chronicle; Greater Houston; Harris County, Texas; Galveston, Texas; Downtown Houston; Houston Rockets; Texas
Seattle	Seattle; Seattle Seahawks; Seattle metropolitan area; Kobe; Microsoft; Downtown Seattle; Light rail; Alki Point
Nashville, Tennessee	Nashville, Tennessee;Belmont University; Frist Center for the Visual Arts; Southeastern Conference; Centennial Park (Nashville); Gaylord Opryland Resort & Convention Center
Pittsburgh, Pennsylvania	Pittsburgh; Midwestern United States; PNC Park; Station Square; Squirrel Hill (Pittsburgh); Giant Eagle; Fort Pitt Tunnel;Pittsburgh Steelers; Luke Ravenstahl; University of Pittsburgh;

Table 5.4: Examples of Local Entities found in tweets

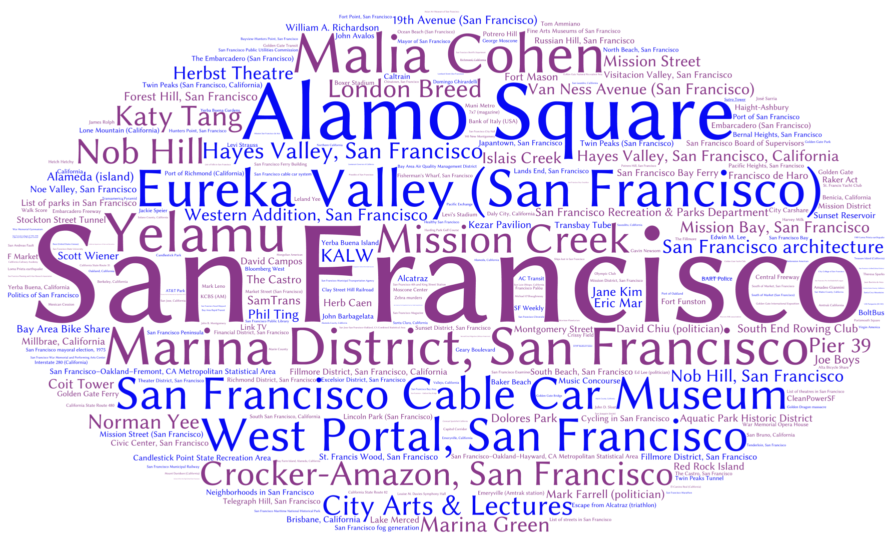

Figure 5.6: Local Entities of San Francisco

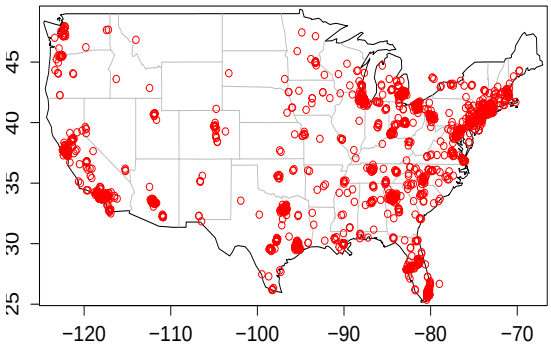

Figure 5.7: Distribution of users predicted within 100 miles of their location

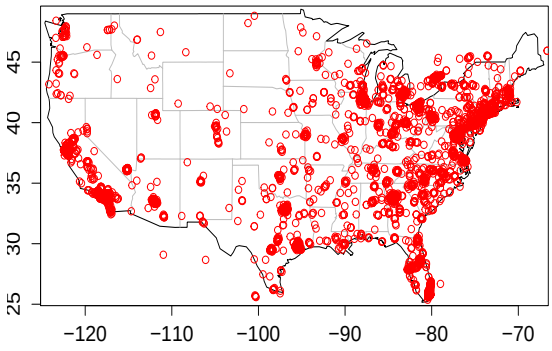

Figure 5.8: Distribution of all users in the dataset

5.3.3 Size of Local Entities

We analyzed the results to understand if the size of the knowledge base, i.e., the number of local entities per city, affect the accuracy of the prediction. The count of local entities in our knowledge base ranges from 11 (for *Island Lake, Illinois*) to 1095 (for *Chicago*). This reflects the information available in Wikipedia about the city. Despite the variation in the amount of information available for each city, we find that our algorithm was able to predict locations of users from 356 distinct cities from our knowledge base having local entities in the range of 40 to 1095. Figure 5.7 shows the distribution of the users, whose location were predicted accurately, across continental United States compared to the distribution of all users in the dataset as shown in Figure 5.8. We see that the accurate prediction (within 100 miles) is not restricted to few cities.

Conclusion and Future Work

In this thesis, we presented a novel knowledge based approach that uses Wikipedia to predict the location of Twitter users. We introduced the concept of *Local Entities* for each city and demonstrated the results of different measures to compute the localness of the entities with respect to a city. Without any training dataset, our approach performs better than the state of the art content based approaches. Furthermore, our approach can expand the knowledge base to include other cities which is remarkably less laborious than creating and modelling a training data set.

In future, we will explore the use of semantic types of the Wikipedia entities to improve the accuracy of the location prediction and decrease the average error distance. We also plan to augment our knowledge base with location information from other knowledge bases such as Geo Names and Wikitravel. Additionally, we will examine how to adapt our approach to predict the location of a user at a finer granularity level like the neighbourhoods in a city.

Bibliography

[1] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging Web Content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280. ACM, 2004.

[3] Thirunarayan Anantharam, Barnaghi and Sheth. Extracting city traffic events from social streams. Technical report, 2014.

[4] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. Semrank: Ranking Complex Relationship Search Results on the Semantic Web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 117–127, New York, NY, USA, 2005. ACM.

[5] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[6] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial Variation in Search Engine Queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[7] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[8] Adam Bermingham and Alan F Smeaton. On using twitter to monitor political sentiment and predict election results. 2011.

[9] Orkut Buyukokkten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting geographical location information of web pages. 1999.

[10] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @ phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. In *ASONAM 2012*, 2012.

[11] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are Where you Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[12] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990.

[13] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommen-

dation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 153–162. ACM, 2012.

[14] Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

[15] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. 2000.

[16] Derek Doran, Swapna Gokhale, and Aldo Dagnino. Accurate local estimation of geo-coordinates for social media posts. In *Proceedings of 26th international conference on Software Engineering and Knowledge Engineering*, pages 642–647.

[17] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[18] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[19] Evgeniy Gabrilovich and Shaul Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI*, volume 6, pages 1301–1306, 2006.

[20] Yegin Genc, Yasuaki Sakamoto, and JeffreyV. Nickerson. Discovering context: Classifying tweets through a semantic transform based on wikipedia. In DylanD. Schmorrow and CaliM. Fidopiastis, editors, *Foundations of Augmented Cognition. Directing*

the Future of Adaptive Systems, volume 6780 of *Lecture Notes in Computer Science*, pages 484–492. Springer Berlin Heidelberg, 2011.

[21] Daniel Gruhl, Meena Nagarajan, Jan Pieper, Christine Robson, and Amit Sheth. *Context and domain knowledge enhanced entity spotting in informal text*. Springer, 2009.

[22] Joseph Hassell, Boanerges Aleman-Meza, and I Budak Arpinar. *Ontology-driven automatic entity disambiguation in unstructured text*. Springer, 2006.

[23] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[24] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. ACM, 2009.

[25] Yuheng Hu, Ajita John, Fei Wang, and Subbarao Kambhampati. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *AAAI*, volume 12, pages 59–65, 2012.

[26] Yuheng Hu, Fei Wang, and Subbarao Kambhampati. Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2640–2646. AAAI Press, 2013.

[27] Elias Iosif and Alexandros Potamianos. Unsupervised semantic similarity computation between terms using web documents. *Knowledge and Data Engineering, IEEE Transactions on*, 22(11):1637–1647, 2010.

[28] Jaap Kamps and Marijn Koolen. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 232–241. ACM, 2009.

[29] Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 99–113. Springer International Publishing, 2014.

[30] Satya Katragadda, Miao Jin, and Vijay Raghavan. An unsupervised approach to identify location based on the content of users tweet history. In *Active Media Technology*, pages 311–323. Springer, 2014.

[31] Saurabh Khanwalkar, Marc Seldin, Amit Srivastava, Anoop Kumar, and Sean Colbath. Content-based Geo-Location Detection for Placing Tweets Pertaining to Trending News on Map. In *The Fourth International Workshop on Mining Ubiquitous and Social Environments*, page 37, 2013.

[32] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm Eating a Sandwich in Glasgow: Modeling Locations With Tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.

[33] Nicolas Kourtellis, Tharaka Alahakoon, Ramanuja Simha, Adriana Iamnitchi, and Rahul Tripathi. Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, 3(4):899–914, 2013.

[34] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *ICWSM*, 2011.

[35] Chenliang Li, Aixin Sun, and Anwitaman Datta. A generalized method for word sense disambiguation based on wikipedia. In *Advances in Information Retrieval*, pages 653–664. Springer, 2011.

[36] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *arXiv preprint arXiv:1403.2345*, 2014.

[37] Alexander Markowetz, Thomas Brinkhoff, and Bernhard Seeger. Exploiting the internet as a geospatial database. In *International Workshop on Next Generation Geospatial Information*, pages 19–21, 2003.

[38] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location Prediction in Social Media Based on Tie Strength. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 459–468. ACM, 2013.

[39] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[40] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose. *Proceedings of ICWSM*, 2013.

[41] Subhabrata Mukherjee and Pushpak Bhattacharyya. Wikisent: weakly supervised sentiment analysis through extractive summarization with wikipedia. In *Machine Learning and Knowledge Discovery in Databases*, pages 774–793. Springer, 2012.

[42] Andrea Giovanni Nuzzolese, Aldo Gangemi, Valentina Presutti, and Paolo Ciancarini. Encyclopedic knowledge patterns from wikipedia links. In *The Semantic Web–ISWC 2011*, pages 520–536. Springer, 2011.

[43] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[44] Pavel Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–61. Citeseer, 2008.

[45] Sujan Perera, Cory Henson, Krishnaprasad Thirunarayan, Amit Sheth, and Suhas Nair. Data driven knowledge acquisition method for domain knowledge enrichment in the healthcare. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–8. IEEE, 2012.

[46] Ray Richardson, A Smeaton, and John Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. Technical report, Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University, 1994.

[47] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[48] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's @wally?: A Classification Approach to Geolocating Users Based on their Social Ties. In *HT*, 2013.

[49] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[50] Martin Szomszor, Patty Kostkova, and Ed De Quincey. # swineflu: Twitter predicts swine flu outbreak in 2009. In *Electronic Healthcare*, pages 18–26. Springer, 2012.

[51] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1397–1405. ACM, 2011.

[52] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska De Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[53] Amos Tversky. Features of Similarity. *Psychological review*, 84(4):327, 1977.

[54] Pu Wang and Carlotta Domeniconi. Towards a Universal Text Classifier: Transfer Learning using Encyclopedic Knowledge. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 435–440. IEEE, 2009.

[55] I Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[56] Arjumand Younus, M Atif Qureshi, Fiza Fatima Asar, Muhammad Azam, Muhammad Saeed, and Nasir Touheed. What do the average twitterers say: A twitter model for public opinion analysis in the face of major political events. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 618–623. IEEE, 2011.